Who Lies on Surveys, and What Can We Do About It?

Satoshi Kanazawa[1]
London School of Economics and Political Science

Focusing on demographic characteristics, the author seeks partially to replicate, with a larger set of variables, Belli, Traugott, and Beckmann's (2001) recent study on vote overreports using the same data from the U.S. National Election Studies (1948-1998). His analyses show that Blacks and residents of the Southern States in general are most likely to make false statements on how they voted. He suggests a possible solution for inaccuracies in survey data and proposes that, when validation of verbal responses is not possible, it may be prudent, if feasible, to re-estimate models with and without Black and Southern respondents to make sure that findings are robust.

Key Words: Center for Political Studies; the National Election Studies; surveys; demography.

Empirical support for a scientific theory depends not only on the validity of the theory but also on the quality of the data. Given that a large number of social scientific studies rely on surveys, the quality of survey data, whether respondents answer the questions accurately, is crucially important for the cumulative knowledge in social sciences. If there are *systematic* (as opposed to random) errors in survey responses, some of what we think we know in social sciences may not be true. Accuracy of responses is not the only determinant of the quality of data, but it is an important one. It is not a sufficient condition for quality data, but it is a necessary one.

Of course, under ordinary circumstances, it is very difficult, if not impossible, to ascertain whether respondents are being truthful in their answers because the true answers to the questions are usually unknowable to the interviewers. The National Election Studies, conducted by the Center for Political Studies at the University of Michigan, provide a unique opportunity to ascertain the truthfulness of survey responses,

[1] Address for correspondence: Satoshi Kanazawa, Interdisciplinary Institute of Management, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom. Email: S.Kanazawa@lse.ac.uk. I thank Richard Lynn for his comments on an earlier draft.

which few other data sets offer. Over the last half-century, they have conducted *validation studies*: The interviewers first ask the respondents whether they are registered to vote in their precinct and whether they voted in the last election, and then they go to the respondent's precinct to look up the registration and voter turnout record to verify the respondent's verbal responses. The National Election Studies therefore allow researchers to determine who are more likely to give truthful responses to survey questions and who are more likely to misreport.

Past analyses of the *demographic* determinants of misreporting have, however, produced contradictory findings, with respect to the effect of sex (Belli, Traugott, and Beckmann 2001; Hill and Hurley 1984; Traugott and Katosh 1979); age (Belli et al. 2001; Hill and Hurley 1984; Katosh and Traugott 1981; Sigelman 1982; Traugott and Katosh 1979; Weiss 1968); education (Belli et al. 2001; Bernstein, Chadha and Montjoy 2001; Hill and Hurley 1984; Silver, Anderson and Abramson 1986; Traugott and Katosh 1979); and income (Hill and Hurley 1984; Katosh and Traugott 1981; Traugott and Katosh 1979). These studies therefore do not allow researchers unequivocally to predict who are most likely to misreport their behavior on survey questions. The only demographic characteristic that consistently and reliably predicts vote misreporting is race: Blacks are more likely to lie than non-Blacks (Anderson, Silver, and Abramson 1988; Bernstein et al. 2001; Belli et al. 2001; Hill and Hurley 1984; Katosh and Traugott 1981; Sigelman 1982; Traugott and Katosh 1979). This conclusion from the National Election Studies is consistent with findings from other analyses of survey responses on entirely different behavior. For instance, Mensch and Kandel (1988) report that Blacks and Hispanics are most likely to misreport their drug use.

Belli et al. (2001) conduct the most comprehensive analysis of factors that predict misreporting of voting in the National Election Studies. They include a host of social (demographic), attitudinal, and contextual factors in an attempt to discriminate overreporters from validated voters and admitted nonvoters. Their very ingenious analysis shows that the effects of demographic variables depend on to which group one compares the misreporters. For instance, both age and education have *negative* effects on misreporting among self-reported voters (in a comparison of misreporters with validated voters), but *positive* effects on misreporting among validated nonvoters (in a comparison of misreporters with admitted nonvoters). Sex has no significant effect among the first group, but a significantly negative effect among the second. Once again, the only demographic variable which consistently predicts misreporting in both comparisons is race: Nonwhites are significantly more likely to lie than whites.

In this brief research note, I will try to replicate Belli et al's (2001) findings in an entirely different specification. I will include a much larger set of social, political and demographic variables in an attempt to predict misreporting of both voter registration and voting (in a comparison of misreporters, on the one hand, with both validated voters and self-admitted nonvoters, on the other). The expanded set of independent variables in the logistic regression will allow me to assess whether Belli et al. (2001) has neglected some other potential predictors of misreporting. I also suggest one possible solution for the problem of misreporting in survey research.

Data

The Center for Political Studies has conducted the National Election Studies since 1948. For each Presidential and Congressional election, a representative sample of adult, non-institutional citizens of the United States are interviewed in person and asked a large number of questions about their political attitudes and behavior. Two of these questions ask whether the respondents are registered to vote in their precinct, and whether they voted in the last election. In eight different years (1964, 1976, 1978, 1980, 1984, 1986, 1988, 1990), the interviewers verify the respondents' verbal responses about their registration and voting, by looking up the precinct records to see if they are really registered and if they really voted in the last election. I pool data from these eight separate years in my analyses below.

Dependent Variables

I use two dependent variables in my statistical analyses: Whether the respondent is accurate or truthful about voter registration, and whether the respondent is accurate or truthful about voting in the last election. In the National Election Studies cumulative data (1948-1998)

(VAR CF9154-9155), the respondent is assumed to be truthful if she states that she is not registered to vote or that she did not vote in the last election. Only if the respondent states that she is registered to vote or voted in the last election is the attempt made to verify her verbal responses. Belli et al. (2001, p. 483, Table 1) report that only 0.0 to 1.4% of respondents in individual surveys underreport their voting, reporting not to have voted when they did. If the precinct record is consistent with the verbal response, then the respondent is coded as truthful in her response (=1); if not, then the respondent is coded as untruthful (=0). Since both dependent variables are binary, I use logistic regression to estimate the effects of individual characteristics on the truthfulness of their survey responses.

Table 1
Truthful Reponses on Registration and Voting

	Registration	Voting
Demographic characteristics		
Race (Black=1)	-.4830**	-.4394**
	(.1740)	(.1581)
Hispanicity (Hispanic=1)	-.3647	.1750
	(.2663)	(.2768)
Age	.0141***	.0052
	(.0039)	(.0033)
Marital status	.2389	.3307**
(Currently married=1)	(.1287)	(.1138)
Sex (Male=1)	.0530	-.0667
	(.1194)	(.1052)
Party affiliation		
(Reference=Independent)		
Democrat	-.2171	-.3458*
	(.1786)	(.1685)
Republican	-.0222	-.2534
	(.1980)	(.1812)
Social class		
Subjective class	.0178	-.0049
	(.0384)	(.0338)
Education	-.0843	-.0807
	(.0487)	(.0425)
Income	.0169	-.0077

	(.0625)	(.0556)
Duncan SEI	.0003	1.01^{-5}
	(.0003)	(.0003)
Religion		
(Reference=None/Other)		
Catholic	-.1270	-.5593**
	(.2174)	(.2092)
Protestant	.0576	-.3341
	(.1968)	(.1929)
Jewish	-.6002	-.3175
	(.3645)	(.3783)
Region		
(Reference=Northeast)		
North Central .	-.0930	-.3043
	(.1905)	(.1683)
South	-.4368*	-.6254***
	(.1768)	(.1606)
West	-.1348	-.2009
	(.2023)	(.1840)
Year	.0953****	.0659**
	(.0232)	(.0201)
Constant	-186.653	-127.442
	(45.9634)	(39.8036)
-2 log likelihood	2205.796	2703.830
c^2 (df = 18)	80.888****	72.863****
% correctly identified	91.10	88.23
n	3,807	3,832

Note: Main entries are unstandardized coefficients, and numbers in parentheses are standard errors.

* $p < .05$ ** $p < .01$ *** $p < .001$ **** $p < .0001$

Independent Variables

Demographic characteristics

I regress each of the dependent variables on a number of demographic characteristics of the respondent: Race (1 if Black; 0 otherwise); Hispanicity (1 if Hispanic; 0 if otherwise); Age; Marital status (1 if currently married; 0 if otherwise); and Sex (1 if male; 0 if female).

Party affiliation

I include two dummies to measure the respondent's party affiliation, Democrat and Republican, with independent as the reference category.

Social class

I enter several variables into the logistic regression to assess the effects of social class: Subjective class (a 7-point scale that measures the respondent's subjective class identity); Education (a 7-point scale that measures the respondent's educational attainment); Income (a 5-point scale constructed by the Center for Political Studies from reported income figures); and Duncan SEI.

Religion

I measure the respondent's religious affiliation with a series of dummies, Catholic, Protestant, and Jewish, with "none/other" as the reference category.

Region

I measure the respondent's geographic location within the United States with a series of dummies, North Central, South, and West, with Northeast as the reference category.

Year

Finally, in order to assess and control for any long-term linear trend in the truthfulness of survey responses or any changes in the validation of registration and voting, I enter the year of the survey in the logistic regression equations.

Results

Table 1 presents the results of the logistic regression analyses. The most salient finding is that individual characteristics tend not to have consistent effects on the two dependent variables. Age has a significantly positive effect on the truthfulness of the verbal response about registration (older respondents are more truthful), but not about voting. Similarly, currently married respondents are more likely to be truthful, and Democrats and Catholics are less likely to be truthful, about their voting, but not about their registration.

There are only two individual characteristics that have a consistent effect on the two dependent variables: Race and South. Relative to non-

Blacks, Blacks are significantly ($ps < .01$) less likely to be truthful about both registration and voting. Similarly, relative to residents in the Northeastern states, Southerners are significantly less likely to be truthful about registration ($p < .05$) and voting ($p < .001$). My analyses therefore partially replicate Belli et al.'s (2001) earlier study. Like them, I find that Blacks are less likely to be truthful in their responses. Region is not included in Belli et al.'s (2001) analysis, nor in many previous analyses of vote validation in the National Election Studies.

The greater tendency of Blacks to lie about their behavior may be attributable to their higher likelihood of having psychopathic personality (Lynn 2002). I am not aware of any potential explanation for the Southerners' greater tendency to misreport their behavior.

The survey year also has a consistently positive effect on the truthfulness of both responses, replicating Belli et al's (2001) study. I am not quite sure what to make of this finding, but it seems somewhat unlikely that Americans have become more honest in their survey responses in the quarter century from 1964 to 1990. One possible interpretation is that the validation techniques have improved over the years, and now the interviewers are better able to validate respondents' verbal responses accurately with the precinct records. If this interpretation is correct, then it bolsters the findings about Blacks and Southerners mentioned above. It indicates that the negative effect of being Black or Southerner is independent of the sophistication of the validation techniques, and that it is *not* because Blacks and Southerners tend to live in precincts where validation is more difficult that these categories of respondents *appear* to be less truthful. Another potential explanation for the positive effect of survey year is that, as voter turnout continues to decline over the years, it has become more socially acceptable not to vote, and nonvoters are now under less pressure to misreport their behavior for social desirability.

A Possible Remedy: An Illustration

My finding that Blacks and Southerners are less truthful in their answers to survey questions must be replicated in future studies, ideally involving responses that have nothing to do with voter registration and voting. If no other data are available that allow researchers to validate the respondents' verbal responses to survey questions, perhaps experimental studies can be designed (as in Rasinski 1999). If, on the other hand, the findings turn out to be robust and generalizable to other survey questions and responses about other types of behavior, then maybe future researchers might consider estimating parameters in their analyses of survey data with and without Blacks or Southerners included in their sample, to see if there are any differences in substantive conclusions drawn. Of course, this option is not available if the main focus of the research is the effect of race or the region of residence.

Table 2 illustrates the potential utility of such an approach. The logistic regression equations predict who turns out to vote at the election from a set of demographic and political characteristics of the respondent. As the first column shows, if the dependent variable is *reported* voting, then it appears that men are significantly ($p < .01$) more likely to vote than women.

The sex of the respondent ceases to have a significant effect on the likelihood of voting, however, if I estimate the model without Black and Southern respondents (Table 2, second column), while the significance of all the other variables remains. Thus, on the basis of logistic regression only with respondents who are most likely to be truthful, I would conclude that, unlike all the other variables included in the model, sex of the respondent has no significant effect on the likelihood of voting, and men and women are equally likely to vote.

Table 2
Illustration: Estimating the Model with and without Black and Southern Respondent

	Reported voting		Validated voting
		Blacks and Southerners excluded	
	Full sample		Full sample
Sex (Male=1)	*.0824***	*.0173*	*-.0875*
	(.0297)	*(.0393)*	*(.0472)*
Age	.0353****	.0369****	.0326****
	(.0010)	(.0013)	(.0015)
Marital status	.3335****	.4408****	.3507****
(Currently	(.0330)	(.0439)	(.0514)
married=1)			

Subjective class	.0412****	.0529****	.0317*
	(.0096)	(.0128)	(.0150)
Education	.3179****	.3053****	.2464****
	(.0106)	(.0143)	(.0169)
Income	.2795****	.2741****	.2051****
	(.0158)	(.0212)	(.0255)
Democrat	.7947****	.7739****	.5393****
	(.0438)	(.0569)	(.0732)
Republican	.8367****	.8592****	.7457****
	(.0468)	(.0591)	(.0774)
Constant	-3.8164	-3.7068	-3.3378
	(.0818)	(.1113)	(.1299)
-2 log likelihood	27920.279	16160.182	10880.685
c^2(df=8)	3983.212****	2230.679****	1144.384****
% correctly identified	71.41	74.76	67.56
n	24,860	15,418	8,983

Note: Main entries are unstandardized coefficients, and numbers in parentheses are standard errors.

* $p < .05$ ** $p < .01$ *** $p < .001$ **** $p < .0001$

The sex of the respondent ceases to have a significant effect on the likelihood of voting, however, if I estimate the model without Black and Southern respondents (Table 2, second column), while the significance of all the other variables remains. Thus, on the basis of logistic regression only with respondents who are most likely to be truthful, I would conclude that, unlike all the other variables included in the model, sex of the respondent has no significant effect on the likelihood of voting, and men and women are equally likely to vote.

This conclusion, in fact, appears to be true. The third column in Table 2 shows that, if the dependent variable is *validated* (not reported) voting, where the respondents cannot misrepresent themselves, then sex has no significant effect on the likelihood of voting. Validation of respondents' verbal responses to survey questions, of course, is usually not possible. In such a case, if the main focus of the research is not to estimate the effect of race or the region of residence, it might be prudent to estimate the model with and without Black and Southern respondents to see if the main findings are robust across samples.

References

Anderson, Barbara A., Brian D. Silver, and Paul R. Abramson
1988 "The Effects of Race of the Interviewer on Measures of Electoral Participation by Blacks in SRC national Election Studies." *Public Opinion Quarterly.* 52: 53-83.

Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann
2001 "What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics.* 17: 479-498.

Bernstein, Robert, Anita Chadha, and Robert Montjoy
2001 "Overreporting Voting: Why It Happens and Why It Matters." *Public Opinion Quarterly.* 65: 22-44.

Hill, Kim Quaile and Patricia A. Hurley
1984 "Nonvoters in Voters' Clothing: The Impact of Voting Behavior Misreporting on Voting Behavior Research." *Social Science Quarterly.* 65: 199-206.

Katosh, John P. and Michael W. Traugott
1981 "The Consequences of Validated and Self-Reported Voting Measures." *Public Opinion Quarterly.* 45: 519-535.

Lynn, Richard
2001 "Racial and Ethnic Differences in Psychopathic Personality." *Personality and Individual Differences.* 32: 273-316.

Mensch, Barbara S. and Denise B. Kandel
1988 "Underreporting of Substance Use in a National Longitudinal Youth Cohort." *Public Opinion Quarterly.* 52: 100-124.

Rasinski, Kenneth A., Gordon B. Willis, Alison K. Baldwin, Wenchi Yeh, and Lisa Lee
1998 "Methods of Data Collection, Perception of Risks and Losses, and Motivation to Give Truthful Answers to Sensitive Survey Questions." *Applied Cognitive Psychology.* 13: 465-484.

Sigelman, Lee
1982 "The Nonvoting Voter in Voting Research." *American Journal of Political Science.* 26: 47-56.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson
1986 "Who Overreports Voting?" *American Political Science Review.* 80: 613-624.

Traugott, Michael W. and John P. Katosh
1979 "Response Validity in Surveys of Voting Behavior." *Public Opinion Quarterly.* 43: 359-377.

Weiss, Carol H.
1968 "Validity of Welfare Mothers' Interview Responses." *Public Opinion Quarterly.* 32: 622-633.